

连锁分析及关联分析

本章介绍连锁分析及关联分析的基本概念、方法及其应用。

本世纪初人们就开始研究连锁现象，然而只是从八十年代开始，大量可供参考的基因标记(genetic marker)的出现和计算机技术的发展才使之有了长足的进展，大量备选基因(candidate gene)的出现也促成了关联分析的研究与应用。

目前连锁分析和关联分析研究的热点是复杂疾病如癌症、多发性硬化症、哮喘病、原发性高血压、糖尿病，一些精神性疾病如精神分裂症、双相性情感障碍等。

一、连锁分析

连锁(linkage)是一种遗传现象，是指当同一染色体上的某些位点由于相距很近，在减数分裂过程中这些位点发生交叉的机率较小，这些位点的等位基因连锁在一起从亲代遗传到子代。位点间的距离越近，在减数分裂过程中发生交叉的可能性越小，位点间重组的可能性越小。在下面的三代家系(图1)中，第三代四兄弟姊妹中，个体Ⅲ1和个体Ⅲ2没有发生重组，而个体Ⅲ3和Ⅲ4发生了重组。

(图1)

位点间遗传距离的大小可用重组分数(recombination fraction)进行估计。重组分数通常用 θ 表示，两个位点间的重组分数定义为： $\theta = \text{重组配子数} / (\text{重组配子数} + \text{非重组配子数})$ ，其取值范围为0——0.5。如果位点位于不同的染色体上，或位于同一染色体上但相距甚远，那么在减数分裂过程中重组的机率趋于50%。遗传连锁图上基因之间的相对距离用摩尔根单位(Morgan unit, 简称M)来表示，表明染色体的一次交叉。1/100摩尔根为一个厘摩(cM)，表示重组值为1/100，大约相当于百万个DNA碱基对。

我们把基因按照其在染色体上的相对位置顺序排列，就得到了基因图(genetic map)。图距与重组分数之间的函数关系称为作图函数(map function)，它们可以用图象形象地表示出来(图2)。常用的作图函数是Haldane和Kosambi函数，其数学公式是： $x = -0.5 \ln(1 - \theta)$ 及 $x = 0.25 \ln \frac{1 + 2\theta}{1 - 2\theta}$ 。它们分别用于多个位点和位点之间没有干扰和有干扰的情形。

(图2)

连锁分析是利用连锁的原理研究致病基因与参考位点的关系。根据孟德尔分离率，如果同一染色体上的位点不连锁，那么基因标记将独立于致病基因而分离，与致病基因位于同一单倍体或不同单倍体的机会各占一半，否则表明连锁的存在。连锁分析通常采用家系分析的方法。

连锁分析可以采用直接方法和极大似然方法。直接法是根据所有减数分裂中重组发生的相对频率计算重组率，其计算比较简单。但由于大多数情况下确定重组与否比较困难，直接法便不适用。极大似然的结果主要用连锁分数(lod score)来表示，它是连锁相对于不连锁的概率比数取常用对数，因此该值为3表明连锁相对于不连锁的可能性为一千倍，我们定义为连锁；反之若该值为-2则排除连锁。这一概念和方法早于本世纪五

十年代提出,适于计算机的算法于七十年代初被提出,适应大量基因标记需要的新算法于八十年代末提出。极大似然法又称为参数方法,因为在进行分析时要指明疾病的遗传方式,如显性、隐性?是否性连锁?主要是疾病的基因频率、外显率,这些参数的来源主要是一些临床和流行病学资料,也可以通过分离分析来得到。由于某些疾病的遗传方式不明确,复杂疾病的外显率较低,人们提出一些非参方法,如发病的同胞对和其他亲属对分析。尽管这些方法也会受到疾病模式的影响,但在分析时不需要指明疾病模式。

进行连锁分析主要借助于计算机软件,其中较常用的是LINKAGE和GENEHUNTER。LINKAGE用于大家系而每个个体只有少量基因标记的分析;GENEHUNTER则用于小家系而每个个体有大量基因标记的情形,它还给出上述的非参分析,由于其非参统计量在不同标记之间趋于低估,因而人们提出了相应的改进方法。构造基因图可用软件CRIMAP,为了在世界范围内的规格化,国际上采用法国人类多态中心(CEPH)的家系作为标准。

二、关联分析

由于同一染色体上位点间紧密连锁或其它原因,在同一配子中某些等位基因的组合可能增加,这种遗传现象称为等位基因的关联(association),也称为连锁不平衡(linkage disequilibrium)。从统计上看,关联发生时单倍型的频率不是组成单倍型的各个位点的基因频率之积。由于关联分析相对于连锁分析的有效性,近来受到更多的重视。

造成连锁不平衡的原因有多种,如:随机遗传漂移、缔造者(founder)效应、突变、选择、不同人群的混合等。一旦在人群演化的某个阶段形成等位基因的关联,由于致病基因与被研究标记间紧密连锁,关联状态可以维持多代(图3)。

(图3)

经典的关联研究(病例对照研究)方法是从人群水平上比较疾病组与对照组之间某个位点基因或基因型频率的差异。但是病例组与对照组的这种频率差异有可能是由于所选择的研究人群的偏倚,如分层和混合等所造成。为了避免这种偏倚,可以用患者双亲未传递的等位基因作为对照,这是单倍型相对危险(HRR)方法的基础。考虑到患者每个父母的标记位点基因型传递与未传递的对应关系,人们又提出了传递不平衡或传递扭曲检验(TDT),它只使用基因位点为杂合子的患者父母,该方法是在发生连锁的前提下检测连锁不平衡的存在,可用于多于两个等位基因的情形。由于一些疾病发病年龄较晚,亲代资料很难收集,近来人们又使用未患病同胞做为对照。

任何列联表分析软件都可以用来比较病例——对照间的频率差异,但若等位基因或基因型分布不集中,患者和对照等位基因或基因型所构成的列联表比较稀疏,需要采用随机模拟、置换或枚举求得精确概率。多个等位基因的罗吉斯蒂(logistic)回归分析可以采用软件ETDT。若果造成

连锁不平衡是由于人群缔造者效应，则使用相应的分析方法更为有效。

连锁分析和关联分析密切相关，一个简单的例子是单独研究一个大家系，如果疾病是常染色体显性遗传，则会容易发现，这种传递不平衡现象的存在，仅仅是由于连锁而致。

如果确认了连锁现象的存在，人们进一步采用物理定位方法，直至确定致病的特定DNA碱基对。此外，大量的生物杂交实验对人类疾病的遗传机制研究也有参考价值。

三、使用连锁分析与关联分析的有关问题

目前，连锁和关联分析研究单基因遗传病比较成功，如Huntington舞蹈症、囊性纤维化。复杂疾病的也有越来越多的报道，如早发乳腺癌和老年痴呆症研究。这些研究成为本领域研究者参照的范例。

连锁分析和关联分析通常从疾病的家族聚集性开始。有效的流行病学研究非常重要，如合适的人群、可靠的疾病人群发病率、不同级亲属间的相对危险度等。大量的多态基因标记，密集的基因图，也是大规模采用连锁分析和关联分析技术进行遗传疾病基因定位的必要前提。

在研究开始时，研究者往往想知道，根据小规模收集的家系，是否能检测到连锁的存在？这要借助所谓的功效分析，可借助计算机软件如SLINK、SIMLINK进行一些模拟分析。另外，近缘亲属遗传定位(homozygosity mapping)方法也常采用。因为人类遗传学研究只能采用被动观测的方法，又由于纯合子的存在，即使亲代和子代信息完整，也无助于连锁的检测，因此信息量有很大的局限性。连锁分析只能检测致病基因与基因标记距离大于百分之一的情形，因此全基因筛选大约需要3000多个位点。关联分析能够大大提高定位的分辨率，结合人群的进化史和单倍型频率分析，分辨可达几十个DNA碱基对。但关联分析需要有效的备选基因。复杂疾病受到许多因素影响，如不完全外显，表现型的模拟，遗传的异质性等，并不服从简单的孟德尔遗传模式。不完全外显是指个体虽然有发病基因但并未发病，或者疾病晚发，从事研究时不能确定是否有基因的携带；表现型的模拟是指病人并不具有致病基因；而位点的遗传异质性是指所研究的家系有的存在连锁现象，而另一部分家系则不然。若疾患诊断标准的掌握不当，统计上重复检验造成的误差等，会造成不同研究重复上的困难。

由于基因检测技术的发达和计算机软件使用日益便利，使大规模的全基因组扫描成为可能。由于重复分析的影响，显著性水准要相应的修正。为免除繁复的基因检测工作，人们不断开发能够进行大量基因检测的技术，如DNA芯片(DNA chips)技术进行“海量”提取、DNA混合分析技术(DNA pooling)是将病例与对照分别混合，从而比较两者之间频率上的差异，藉此对筛选的基因在个体水平上进行检测。

为了适应未来检测的需要，应当建立样品库。为了避免人为误差和有效地使用数据，要进行计算机管理。国际互联网络的采用，加速了信息的传递速度，提高了使用效率。美国国家生物信息中心、国立卫生研究所、英国人类基因工程中心等收集了大量数据和分析软件。发达国家的一些

管理方法, 值得我们借鉴使用。著名的杂志有《细胞》、《自然遗传》、《美国人类遗传学杂志》、《自然》、《科学》、《人类分子生物学》等。本章最后也给出了一些连锁分析和关联分析的参考书目。

连锁分析可用于定位发病基因, 遗传咨询、医学鉴定等。如果结合人群演变史, 那么将会使基因的定位大为精确。

参考书目

Khoury MJ, Beaty TH and Cohen BH. The Fundamentals of Genetic Epidemiology. Oxford University Press, 1993.

Ott J. Analysis of Human Genetic Linkage. The Johns Hopkins University Press, 1991.

Sham P. Statistics in Human Genetics. Edward Arnold, 1998.

Terwilliger JD and Ott J. Handbook of Human Genetic Linkage. The Johns Hopkins University Press, 1994.

Weir BS. Genetic Data Analysis II. Sinauer Associates, 1996.

(赵京华、李涛、沈伯松[Pak C Sham])